

# NAVAL POSTGRADUATE SCHOOL MONTEREY, CALIFORNIA



## THESIS

ANALYSIS OF RUSSIAN AND SPANISH SUBSKILL  
TESTING AT THE DEFENSE LANGUAGE  
INSTITUTE

by

Carlton L. Lavinder III

September 1996

Thesis Advisor:

Lyn R. Whitaker

Approved for public release; distribution is unlimited.

19961220 110

DTIC QUALITY INSPECTED 1

REPORT DOCUMENTATION PAGE			Form Approved OMB No. 0704-0188	
Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instruction, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188) Washington DC 20503.				
1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE September 1996.	3. REPORT TYPE AND DATES COVERED Master's Thesis		
4. TITLE AND SUBTITLE: ANALYSIS OF RUSSIAN AND SPANISH SUBSKILL TESTING AT THE DEFENSE LANGUAGE INSTITUTE		5. FUNDING NUMBERS		
6. AUTHOR(S) Carlton L. Lavinder III				
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Naval Postgraduate School Monterey CA 93943-5000		8. PERFORMING ORGANIZATION REPORT NUMBER		
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)		10. SPONSORING/MONITORING AGENCY REPORT NUMBER		
11. SUPPLEMENTARY NOTES The views expressed in this thesis are those of the author and do not reflect the official policy or position of the Department of Defense or the U.S. Government.				
12a. DISTRIBUTION/AVAILABILITY STATEMENT Approved for public release; distribution is unlimited.			12b. DISTRIBUTION CODE	
13. ABSTRACT (maximum 200 words) The Defense Language Institute is responsible for training military and government service personnel requiring a foreign language skill. Ten Subskill tests have been developed to evaluate the graduating students' language abilities and to determine if they have met the sponsor's Final Learning Objectives. The Subskill tests in some languages have been in place long enough that they can now be studied. This thesis examines these Subskill tests for both Russian and Spanish to determine if the tests have been developed and implemented in a manner to efficiently and consistently discriminate between students of different abilities. Three different issues are treated. The ANOVA is used identify Subskill tests with significant rater effects and the magnitude of those effects when they are present. Item Response Theory is used to examine the Subskill tests at the question level in order to identify questions that poorly discriminate between students of different abilities. In addition, the ability range that students are tested over is examined. Finally, methods using principle components and multiple regression are used to determine which tests, if any, can be eliminated with an acceptable loss of information about the students.				
14. SUBJECT TERMS Inter-rater Reliability, Friedman Non-parametric ANOVA, Item Response Theory.			15. NUMBER OF PAGES 66	
			16. PRICE CODE	
17. SECURITY CLASSIFICATION OF REPORT Unclassified	18. SECURITY CLASSIFICATION OF THIS PAGE Unclassified	19. SECURITY CLASSIFICATION OF ABSTRACT Unclassified	20. LIMITATION OF ABSTRACT UL	

NSN 7540-01-280-5500

Standard Form 298 (Rev. 2-89)  
Prescribed by ANSI Std. Z39-18 298-102



Approved for public release; distribution is unlimited.

**ANALYSIS OF RUSSIAN AND SPANISH SUBSKILL TESTING AT THE  
DEFENSE LANGUAGE INSTITUTE**

Carlton L. Lavinder III  
Lieutenant, United States Navy  
B.S., North Carolina State University, 1989

Submitted in partial fulfillment  
of the requirements for the degree of

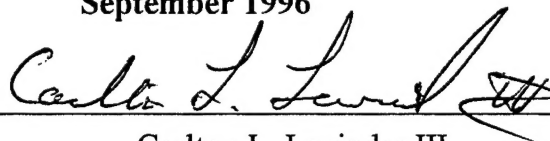
**MASTER OF SCIENCE IN OPERATIONS RESEARCH**

from the


**NAVAL POSTGRADUATE SCHOOL**

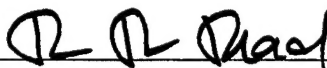
September 1996

Author:

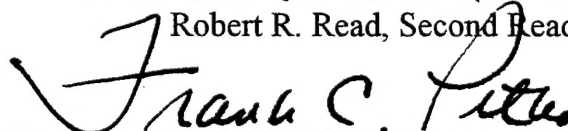
  
Carlton L. Lavinder III

Approved by:

  
Lyn R. Whitaker, Thesis Advisor



Robert R. Read, Second Reader



Frank C. Petho, Chairman

Department of Operations Research



## ABSTRACT

The Defense Language Institute is responsible for training military and government service personnel requiring a foreign language skill. Ten Subskill tests have been developed to evaluate the graduating students' language abilities and to determine if they have met the sponsor's Final Learning Objectives. The Subskill tests in some languages have been in place long enough that they can now be studied. This thesis examines these Subskill tests for both Russian and Spanish to determine if the tests have been developed and implemented in a manner to efficiently and consistently discriminate between students of different abilities. Three different issues are treated. The ANOVA is used identify Subskill tests with significant rater effects and the magnitude of those effects when they are present. Item Response Theory is used to examine the Subskill tests at the question level in order to identify questions that poorly discriminate between students of different abilities. In addition, the ability range that students are tested over is examined. Finally, methods using principle components and multiple regression are used to determine which tests, if any, can be eliminated with an acceptable loss of information about the students.



## TABLE OF CONTENTS

I.	INTRODUCTION .....	1
A.	BACKGROUND .....	1
B.	AREA OF RESEARCH .....	2
1.	Grading Consistency .....	2
2.	Ability Range Tested .....	3
3.	Redundancy of Tests .....	4
C.	OVERVIEW .....	4
II.	DATA .....	5
A.	TESTS/TESTING .....	5
B.	GRADING .....	6
C.	DATA COLLECTION .....	6
1.	Language and Group Selection .....	6
2.	Grading Consistency .....	7
3.	Ability Range Tested .....	7
4.	Redundancy of Tests .....	7
III.	INTER-RATER RELIABILITY .....	9
A.	RELIABILITY OF RATERS .....	9
B.	ANALYSIS .....	9
1.	Non-parametric ANOVA .....	9
2.	Two-way ANOVA .....	11
C.	DISCUSSION OF RESULTS .....	13
IV.	SUBSKILL TEST RESULTS ANALYSIS .....	15
A.	COMPOSITE TEST SCORE ANALYSIS .....	15
B.	ITEM RESPONSE THEORY .....	19
1.	Item Characteristic Curve .....	19
2.	Types of Models .....	20



C.	FITTING THE MODEL .....	22
1.	Model Used .....	22
2.	Item Parameter Estimation .....	23
3.	Student Ability Parameter Estimation .....	24
4.	Preparing the Data .....	25
D.	MODEL FIT .....	25
1.	Screening Results .....	25
2.	Results of IRT Parameter Estimation .....	26
3.	Goodness of Fit .....	27
E.	INTERPRETING THE MODEL .....	29
1.	Information Functions .....	29
2.	Actual Information Functions .....	31
F.	DISCUSSION OF RESULTS .....	32
V.	TEST REDUCTION .....	35
A.	MOTIVATION .....	35
1.	Background .....	35
2.	Methods to be Employed .....	35
3.	Data Examined .....	36
B.	METHODOLOGY .....	36
1.	The Principle Component Method .....	36
2.	The Multiple Correlation Method .....	37
3.	Comparing Methods .....	38
C.	RESULTS .....	38
D.	DISCUSSION OF RESULTS .....	41
VI.	SUMMARY/RECOMMENDATIONS .....	43
A.	SUMMARY .....	43
B.	RECOMMENDATIONS .....	44
	APPENDIX A .....	47

APPENDIX B .....	49
LIST OF REFERENCES .....	51
INITIAL DISTRIBUTION LIST .....	53



## EXECUTIVE SUMMARY

The Defense Language Institute is responsible for training military and government service personnel requiring a foreign language skill. Ten Subskill tests have been developed to evaluate the graduating students' language abilities and to determine if they have met the sponsor's Final Learning Objectives. The Subskill tests in Russian and Spanish have been in place long enough that it is now possible to evaluate them and determine if any changes should be made to improve their consistency and efficiency. Three different issues are treated. The consistency of Subskill test grading between raters, whether the tests can distinguish between students of different abilities, and the degree of redundancy of the battery of Subskill tests.

An analysis of variance shows an inconsistency between the raters score assignments for the majority of Subskill tests in Russian and Spanish. In particular, the Spanish FLO 30 and FLO 90 have the largest magnitude of rater effects. This lack of consistency affects the ability of DLI to compare students whose tests were not graded by the same rater.

An efficient test is made up of questions that discriminate between students of different ability and are of different difficulty levels. The application of Item Response Theory to the Spanish Subskill test shows that Spanish FLO's 30 and 40 consist of questions that discriminate between students of different abilities. In addition, these tests test over a wide range of ability. In contrast, the Spanish FLO's 60, 70, 80, and 90 consist of questions that do not discriminate between students of different ability levels, nor do they test over a wide range of ability. The questions that do not discriminate create unnecessary variance in the data and provide no information about the student's abilities.

Through the use of multiple correlation and principle components, it is determined that the Spanish FLO 40 and Russian FLO 30 can be removed from the battery of tests given to graduating students with less than a 5 percent loss of variance in the data about the students. This will allow for savings in the cost of administering the tests, including

the cost of grading the tests as well as a reduction in the time required for the students to take the tests.

Each of the above findings should be addressed to ensure that the Subskill tests provide consistent, efficient data to be used to compare students and determine if students have met the training objectives. A change in any one of the three areas will effect the remaining two. It is recommended that the grading inconsistency be corrected first and the removal or elimination of any test be the last of the three changes. It will be necessary to record the students' scores on individual questions to make these changes. By recording this information DLI will be able to better monitor the Subskill tests in the future.

## **I. INTRODUCTION**

### **A. BACKGROUND**

The Defense Language Institute (DLI) is responsible for training military and government service personnel requiring a foreign language skill. The National Security Agency (NSA) and the Defense Intelligence Agency (DIA) set the standards for the vast majority of students in the Defense Foreign Language Program. In the early 1990's these two communities developed specific training objectives for students entering professional fields in intelligence. DLI was able to combine the requirements from both communities into a single set of program objectives for all students. These program objectives are referred to as Final Learning Objectives or FLO's. Subskill tests were developed by DLI to be used with the Defense Language Proficiency Tests (DLPT's) to evaluate whether or not the graduating students have met these objectives. (DLI, 1995) These Subskill tests are referred to as FLO 10, FLO 20, ..., FLO 100. The DLPT's have been used since 1958 to evaluate military personnel's language proficiency. Military personnel are given the DLPT's prior to graduation and throughout their careers. The results of the DLPT's are used to award incentive pay for those in billets requiring language skills to ensure that they remain proficient.

Because Subskill tests are much newer than the DLPT's, they have not yet been evaluated. This thesis provides the first such evaluation. It will focus on three issues: the consistency of grading of Subskill tests among raters, whether the tests can distinguish between students of differing abilities, and the degree of redundancy in the combined FLO DLPT battery of tests.

## B. AREA OF RESEARCH

### 1. Grading Consistency

For a test to be useful as a comparison tool it is necessary for the grading to remain consistent without regard to who graded it. In six of the ten Subskill tests, the students respond in English. These tests are graded at the Test Management Center by any one of three GS-5's employed as raters. The remaining four tests are graded at the specific language school. Raters use an answer key to grade these tests, interpreting the correctness of the student's response. Normally, only one rater grades a test, which can lead to different scores depending on which rater graded the test. For this study, to determine if there was a rater effect, all three raters independently graded each student's tests. Over a period of a month, each rater was given all the English response tests to grade. The raters do not make marks on the actual answer sheet so it was possible to ensure that the raters did not know that the study was being conducted. The result of this data collection was one computer scoring sheet from each rater for each student's test. Figure 1.1 shows an example from the data of 10 students' test scores assigned by

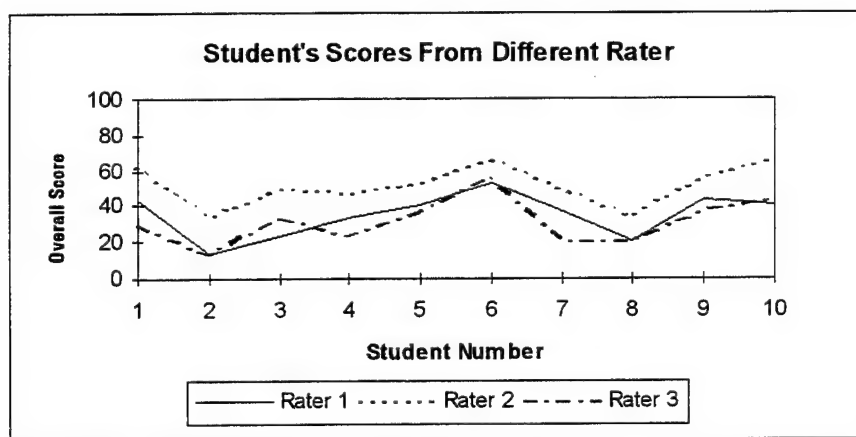


Figure 1.1. An example of difference in student's grades.

different raters from one of the Subskill tests. From this figure it is clear that for these students and for this Subskill test, rater 2's scores are consistently higher than the other two. In fact, rater 2 scored student number seven 30 points higher than rater 3.

## **2. Ability Range Tested**

The purpose of giving the Subskill tests is to determine if the students have met the Final Learning Objectives. To do this it is necessary for the tests to differentiate between students of different abilities and to test over a wide range of abilities. In classical test theory, where composite scores are used, no consideration is given to the difficulty or discriminating power of a test question when grading a test. In Item Response Theory (IRT) each test question can be evaluated for its difficulty and discriminating power, allowing the test developer to determine how much information is provided by each question about the student's ability. Using this information, a test can be constructed to test over a wide range of ability or to ensure that the students meet a certain cut-off ability level. It also allows the test developer to eliminate questions that provide redundant information or no information at all and reduce the length of the test. (Hambleton et al, 1991)

Subskill tests are evaluated using IRT to determine how much information is provided about the student's ability. This evaluation shows that two of the Spanish Subskill tests consist of questions that do differentiate between students and test over a wide range of ability. However, the remaining 4 Spanish Subskill tests graded at the Test Management Center mainly consist of questions that nearly all students get correct or questions that have no discrimination. The result is that these tests do not differentiate between students nor do they test over a wide range of ability. This evaluation will not examine the validity of the questions.



### **3. Redundancy of Tests**

Thirteen tests are currently given to students just prior to graduation. A reduction in the number of tests given while maintaining nearly the same amount of information provided by all the tests would be beneficial in terms of students' time and DLI's budget. Jolliffe (1972,1973) discusses methods of selecting variables (or tests) to remove from data sets while still maintaining nearly the same amount of variance. Two of these methods employing principle components and multiple correlation are used to show that for both Spanish and Russian, one test can be removed with less than a 5 percent loss of total variance in the data set.

### **C. OVERVIEW**

A brief discussion of the testing and grading procedures as well as the collection of data is given in Chapter II. Chapter III addresses inter-rater reliability. In Chapter IV, IRT is used to investigate what range of ability students are tested over on 6 of the 10 Subskill tests. The results of this chapter will allow DLI to decide if test questions should be added or removed to meet their testing objectives. Chapter V examines whether the removal or elimination of one or more FLO's administered just prior to graduation will result in a significant loss of information about the student's ability. Specific recommendations are given in the final chapter.

## II. DATA

### A. TESTS/TESTING

Just prior to graduation, students are given a battery of 13 tests to determine their language proficiency and whether they have met the Final Learning Objectives (FLO's). The DLPT's are used to determine language proficiency in listening, speaking and reading. Subskill tests are used to measure the student's ability to perform the FLO's in the target language and are referred to as FLO 10, FLO 20, ..., FLO 100. The test questions are different for each language. Table 2.1 lists all the tests and gives a brief description of each.

Test	Description
DLPT Listening (List)	Listen to the Target Language
DLPT Reading (Read)	Read the Target Language
DLPT Speaking (Speak)	Speak the Target Language
FLO 10	Elicit Biographical Data (Speaking and Listening)
FLO 20	Two-way Interpretation (Speaking and Listening)
FLO 30 *	Listening: Summarize
FLO 40 *	Listening: Answer Questions
FLO 50	Passage Transcription
FLO 60 *	Number Transcription
FLO 70 *	Reading: Printed Texts
FLO 80 *	Reading: Handwritten Texts
FLO 90 *	Translation: Target Language into English
FLO 100	Translation: English into Target Language

**Table 2.1.** List of tests and a brief description. Tests with an \* are graded at the Test Management Center.

## **B. GRADING**

The tests at DLI are either oral or written response tests. The DLPT speaking, FLO 10, and FLO 20 tests are oral response tests administered and graded at the target language school by a speaker of the target language. DLPT listening and reading are multiple choice tests graded by computer. FLO 50 and FLO 100 are short answer tests whose responses are also in the target language and graded at the target language school. The remaining tests, FLO's 30, 40, 60, 70, 80 and 90, are short answer tests with responses in English and are graded at the Test Management Center independent of the target language. At the time of this study there were 3 employees who graded the tests at the Test Management Center. Raters use only the answer key or protocol, they do not have a copy of the material presented to the students or of the questions. Since the tests are short answer, determining if a response is correct is subjective. The grader fills out a computer scan sheet, recording a 1 for a correct response and 0 for an incorrect response. Each student's test is graded only once. The computer scan sheet and student's answer sheet are stored for a maximum of 3 months and then destroyed due to storage constraints.

## **C. DATA COLLECTION**

### **1. Language and Group Selection**

At the beginning of the study DLI presented a list prioritizing the languages that had classes graduating in the period between January 1996 and March 1996. January was the earliest that data could be gathered once the study was approved and March was picked as the end of data collection to ensure sufficient time to conduct the study. Spanish and Russian were among the higher priority languages and were picked to be studied since classes in both of these languages were graduating in a 3 month time period.

For these two languages, all students who graduated during the time period are included in the inter-rater reliability study and the Item Response Theory study. No other students can be used for those analyses since there is no record of their individual responses to specific questions on each Subskill test.

## **2. Grading Consistency**

In order to compare graders' evaluations of the students' responses, it is necessary to have each rater grade each test. Due to timing and funding limitations, data was gathered on Spanish students tested in January and Russian students tested in February 1996. The Subskill tests that were studied included FLO 30, 40, 60, 70, 80 and 90. The answers of each student were graded separately by each of the three raters. The raters were not informed of the study until after the data was collected. The results of the data collection were three test scores for each of 56 Spanish students and 29 Russian students.

## **3. Ability Range Tested**

The IRT study requires the students' responses to individual questions. Since the score sheets are destroyed after 3 months, the data from the grading consistency study is used. Also for this study the grades from only one grader are used in order to minimize differences in grading criteria. The scores from the most experienced grader are used for this portion of the analysis. Again only FLO 30, 40, 60, 70, 80 and 90 were studied.

## **4. Redundancy of Tests**

A larger data set is needed to examine the redundancy of DLPT's and FLO tests. Data was extracted from the DLI data base on all students graduating from the Spanish and Russian schools October 1994 to March 1996. This data set includes scores on all

tests listed in Table 1. Test scores older than October 1994 are from a different version of the DLPT than is currently used and therefore are not included in the analysis. The data used in this portion of the analysis consists of the 426 Spanish students and 262 Russian students. The original data set contained 529 Spanish and 349 Russian students, but due to missing tests scores, many students in the data set can not have their scores used for the analysis. The large number of missing test scores is attributed to students going to their next duty assignment prior to taking all of the tests. In addition, if students miss a Subskill test for another reason, there is no strong requirement for them to make it up, and no data is available for these students. Since nothing is known about the students who do not take the tests, the results of the analysis can only be applied to the students who do take the test and not the general population.

### **III. INTER-RATER RELIABILITY**

#### **A. RELIABILITY OF RATERS**

For scores on a test to be a useful comparison tool, it is necessary for grading to be consistent, independent of who graded the test. One way that scores from different raters may vary is in the rater's severity. Some raters may tend to give higher scores while others may tend to give lower scores (Longford, 1993). Unless the same rater evaluates all students, there is a possibility that some of the students will receive positively or negatively influenced scores due to the fact that they were graded by a relatively lenient or harsh rater (Raymond, 1990).

Since DLI currently employs 3 raters, any one of which can grade a student's test, it is necessary to ensure that there is no inter-rater reliability problem. In the data gathered to study inter-rater reliability, each student's tests were graded by all three raters. To account for the effect of student ability, students are used as a blocking factor in the analysis. First a nonparametric ANOVA method, the Friedman test, is applied to determine if there is a rater effect for each of the FLO's. Once it has been determined that there is a rater effect, the size of this effect is estimated using Two-way ANOVA.

#### **B. ANALYSIS**

##### **1. Non-parametric ANOVA**

###### **a. Methodology**

As mentioned in Chapter II, the data for the first part of the analysis is in the form of one observation per rater for each student. This format of data is well suited to the Friedman's Test for a randomized block experiment. The hypotheses of interest for

this test is that there is no rater effect. The data is put into a  $J \times 3$  matrix with each column representing a rater or treatment effect  $i, i=1,2,3$ , and the rows represent each student  $j, j=1,\dots,J$ . The scores are ranked across rows with ties receiving midranks. Let  $R_i$  be the sum of the ranks for column  $i$  then the test statistic for the Friedman test is

$$F_r = \frac{12}{IJ(I+1)} \sum_{i=1}^I R_i - 3J(I+1) \quad (3.1)$$

where  $I=3$ .

Under the null hypothesis that there is no rater effect, the test statistic  $F_r$  has approximately a chi-squared distribution with  $I-1$  degrees of freedom. If the resulting p-value is small enough, the null hypothesis, that there is no difference between raters, can be rejected. (DeVore, 1995)

#### **b. Results for Friedman Test**

The results of the Friedman test are shown in Table 3.1. As can be seen

Test	Spanish	Russian
FLO 30	0.0000	0.0000
FLO 40	0.0004	0.0035
FLO 60	0.3620	0.3385
FLO 70	0.0000	0.0492
FLO 80	0.0000	0.0000
FLO 90	0.0000	0.0073

**Table 3.1.** P-values for the Friedman test for a randomized block experiment. P-values less than 0.05 indicate a significant rater effect for that test.

from the table, FLO 60 in both languages are the only tests that have no significant rater effect. The Russian FLO 70 has a higher p-values (0.0492) than the remaining tests, but with a significance value of 0.05 the null hypothesis of no rater effect can be rejected. All of the remaining tests in both languages have p-values small enough to reject the null hypothesis.

## **2. Two-Way ANOVA**

### **a. Methodology**

A two-way additive ANOVA model is used to estimate the magnitude of the rater effects. The factors in the model are the rater and the blocking factor for student effects. Since there is only one observation per cell, it is necessary to assume that there is no interaction between the students and the raters. This presumes, for example, that one rater isn't more lenient with poor students than another. If an interaction term is included in the model, the model would be overparameterized. This analysis is still useful since the presence of a nonzero interaction only reduces the probability that the test will be significant for rater effects (Lindman, R., 1992).

### **b. Results of Two-Way ANOVA Model**

The results of the Two-Way ANOVA model are shown in Table 3.2. As with the Friedman ANOVA, a p-value less than 0.05 allows for the rejection of the null hypothesis that there is no rater effect. These results are consistent with those in the previous section and show that there is a significant rater effect for all tests except the FLO 60's. Examination of the residuals by rater and as a function of the fitted scores support the usual ANOVA assumptions of Normality and equal variance. Further,



Test	Spanish	Russian
FLO 30	0.0000	0.0000
FLO 40	0.0001	0.0019
FLO 60	0.4318	0.2697
FLO 70	0.0000	0.03132
FLO 80	0.0000	0.0000
FLO 90	0.0000	0.0010

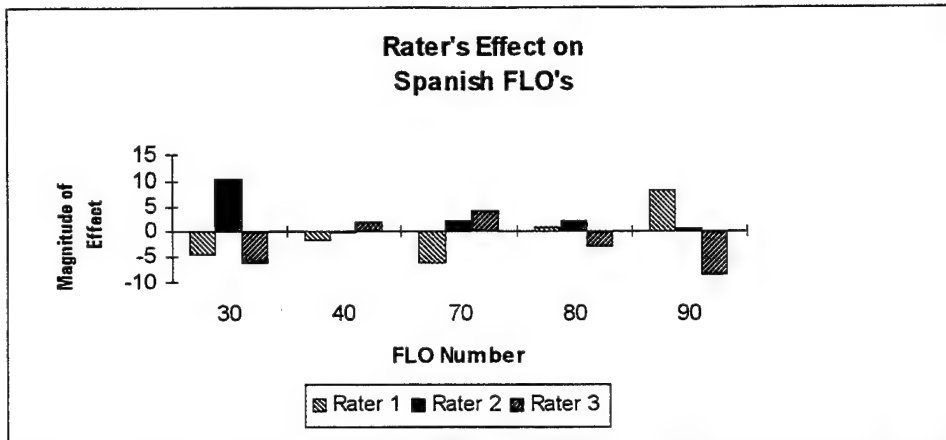
**Table 3.2.** P-values for two-way ANOVA. Values less than 0.05 indicate significant rater effect.

plots of residuals versus fitted values for each rater do not indicate that there is interaction between students and rater.

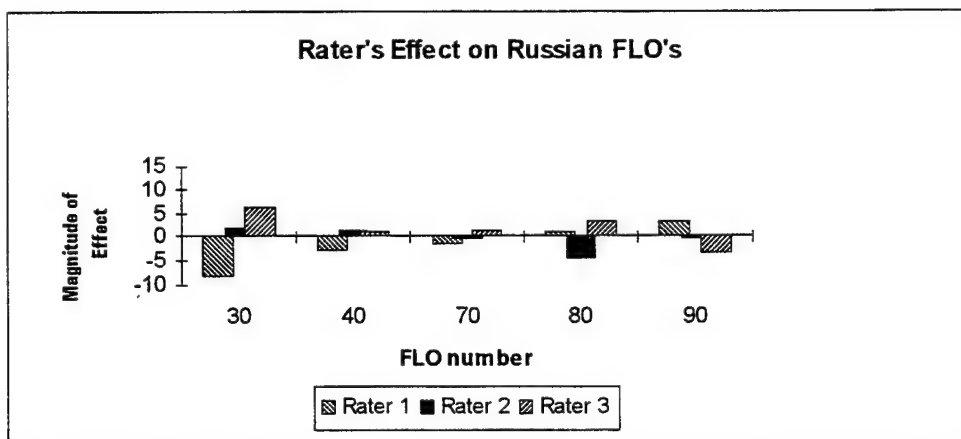
Now that it has been determined that the majority of the tests have significant rater effects, the size of these effects can be estimated. In this model, the effects are parameterized so that

$$\alpha_i = E[\bar{X}_{i.}] - E[\bar{X}_{..}] \quad i=1,2,3 \quad (3.2)$$

where  $\bar{X}_{i.}$  is the average score for the  $i^{th}$  rater and  $\bar{X}_{..}$  is the grand mean. For example, a +10 effect indicates that the rater's expected grades are 10 points more lenient compared to the expected grade averaged over all three raters. The results of the Two-Way ANOVA Model are shown in Figures 3.1 and 3.2 (Appendix A contains the actual values for the effects and standard deviation of each rater on each FLO). From these two figures it is obvious that the Spanish FLO's 30 and 90 have the largest rater effect and should be investigated first. The magnitude of the rater effects on the Spanish FLO 80 and Russian FLO 70 are small and may even be acceptable to the Test Management Center.



**Figure 3.1.** Rater's effect on Spanish FLO's. A +10 effect indicates that the rater grades 10 points more leniently on a test compared to a rater with no effect.



**Figure 3.2.** Rater's effect on Russian FLO's. A +10 effect indicates that the rater grades 10 points more leniently on a test compared to a rater with no effect.

### C. DISCUSSION OF RESULTS

From both the Friedman and Two-Way ANOVA tests, the FLO 60's are the only tests not having a significant rater effect. FLO 60's are the only tests with numbers for answers, which requires little if no subjective interpretation on the part of the raters. Recall that the raters only have the answer key or "protocol" to determine if an answer is correct or incorrect. Unlike FLO 60, the remaining Subskill tests have answers consisting of words and sentences. The two tests with the largest rater effects are Spanish FLO's 30

and 90. The FLO 30 requires the student to summarize what they have heard and the FLO 90 requires the student to translate from the target language into English. Both of these tests seem to require the student to make some decision as to what is important which may be different from the protocol. The remaining tests ask questions or require translation from written texts. These tests may give the student more direction toward a correct answer or what the protocol is looking for. The tests are different for each language which explains why the Russian and Spanish effects are different.

For the tests with effects that DLI finds the Test Management Center can either attempt to compensate for the effects or eliminate them. To compensate for the effects, it would be necessary to record which rater graded each student's test and add or subtract the rater's effect from the student's score. This is possible since there is a block on the computer scoring sheet to indicate which rater graded the test. The method would be effective as long as there is no change in the raters or a change in a rater's effect. Since this is unlikely, it is more useful and effective to reduce or get rid of the rater effects.

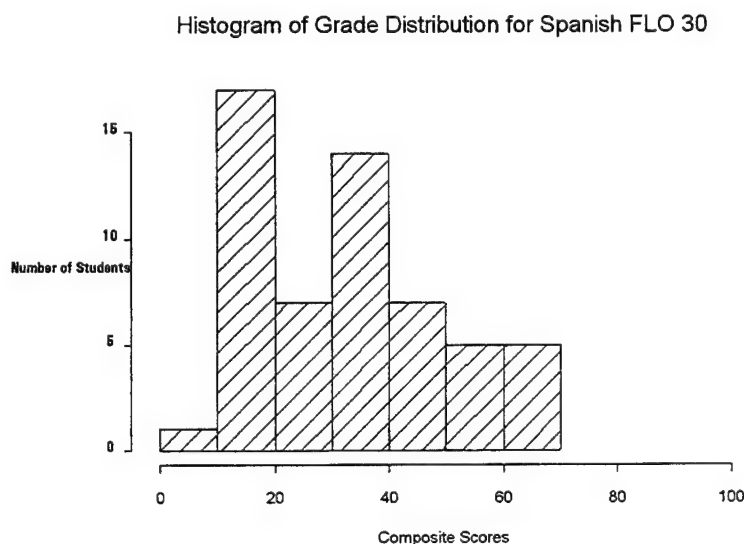
One step in reducing the rater effect is to identify questions whose answers are not well defined and require too much subjective evaluation. Once these questions are identified, the answer key could be rewritten with more specific guidelines to ensure consistent evaluation of the correctness of the answer. Another way to reduce the rater effect may be to include the translation of what the student is presented with as well as the question itself. By allowing the rater to have this information, he should be able to better decide if the student understood the material. In addition, each question could be graded on a scale instead of a 0 or a 1. This would allow the rater to give partial credit to a student who understood the main idea but could not answer the full question.

Finally, the raters should undergo periodic training to reinforce proper grading criteria. The implementation of the suggestions will increase the time required to grade the tests but will create a more stable data base which can be used for future analysis.

## IV. SUBSKILL TEST RESULTS ANALYSIS

### A. COMPOSITE TEST SCORE ANALYSIS

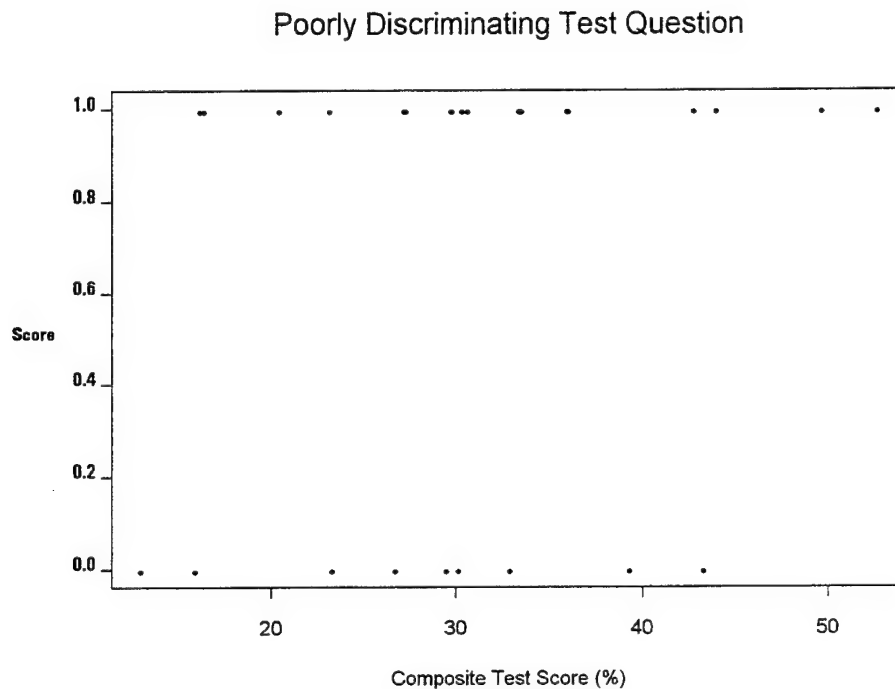
In general, the only data that is recorded at DLI from each of the Subskill tests is the composite score. Normally no record is kept in the data base of the individual question scores. Since each question receives the same number of points for a correct response, there is no difference in points awarded for difficult questions compared to easy questions. A student who answers a difficult question correctly but misses an easy question receives the same score as a student of lower ability who answers the difficult question incorrectly but gets the easy question correct. By keeping only composite scores, DLI has no way of knowing if the Subskill tests are differentiating between students of different abilities. Figure 4.1 shows the histogram of the grades on the Spanish



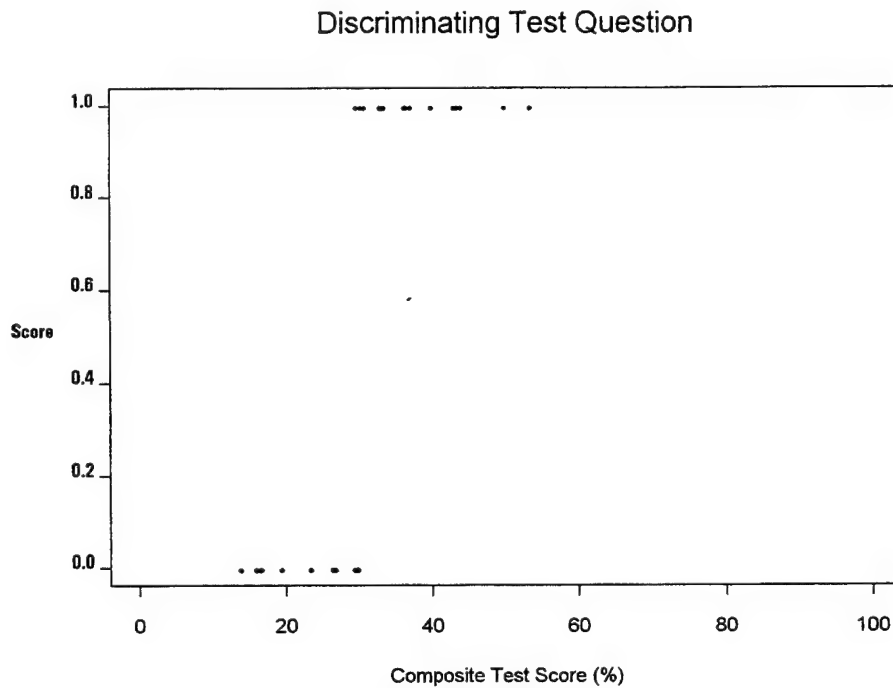
**Figure 4.1.** Histogram of Composite Scores for Spanish FLO 30. Score is in percent.

FLO 30 (To reduce rater effect, all scores used in this chapter are from rater 1). By examining this figure it is impossible to determine whether the test measures over a wide

range of ability and is correctly discriminating between students of different ability or if the test is poorly worded and the students all have the same ability. It is therefore necessary to examine the questions that make up each test for their individual difficulty and how well they discriminate between students of different abilities. A question discriminates well if it is useful for separating students into different ability levels. In Figures 4.2 and 4.3, the scores for two questions (1 for correct and 0 for incorrect) is plotted against the composite test score for each student. Figure 4.2 shows a question with poor discrimination and Figure 4.3 shows a question with good discrimination.



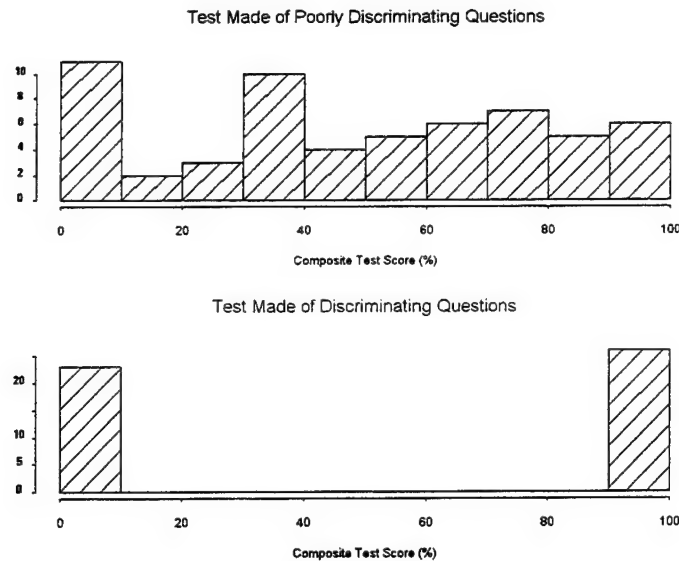
**Figure 4.2.** A test question with poor discrimination.



**Figure 4.3.** A test question with good discrimination.

From Figure 4.2 it can be seen that a student's composite score is not related to his score on the question and thus the question cannot discriminate between students of high and low ability. In contrast, a student's composite score is directly related to his question score in Figure 4.3. This question discriminates well between two levels of student ability.

Two hypothetical tests further show that it is difficult to determine if a test is made of discriminating questions by examining only the composite scores. Figure 4.4 shows histograms of the two hypothetical tests, one made entirely of poorly discriminating



**Figure 4.4.** Histograms of two hypothetical tests, one made of all poorly discriminating questions and the other of all discriminating questions. The y-axis in both cases is the number of students in each bin or group.

questions and the other entirely of question that discriminate well. Scores are evenly distributed for the first test, which may lead an evaluator to incorrectly believe that the test is made of questions that test across the entire ability range of the class. Although the second test sharply divides students into two groups, it provides no further information about a students' ability. By mixing these two types of questions and changing the difficulty level of each question, it is possible to construct a test whose histogram has nearly any shape.

Item Response Theory (IRT) examines the characteristics of the questions that make up the test rather than just the total score. By determining the difficulty and discrimination ability of each question, a test can be constructed to evaluate students over a desired ability range or to ensure that a certain cutoff ability is attained by including questions of a variety of difficulty. By estimating the discrimination factors, questions that are poorly constructed can be eliminated from the test to ensure that the student's true ability is measured.

## B. ITEM RESPONSE THEORY

### 1. Item Characteristic Curve

“Item response theory rests on two basic postulates: (a) The performance of an examinee on a test item(question) can be predicted by a set of factors called ... and (b) the relationship between examinees' item (question) performance and the set of traits underlying item performance can be described by a monotonically increasing function called an item characteristic curve (ICC)” (Hambleton et al,1991). In other words, if the trait used to predict the probability of the student getting a correct response is the student's ability to perform a task, then as the ability of the examinee increases, the probability of responding correctly to the question increases. The ICC is defined by plotting the proportion of correct responses (or probability of answering correct) for each ability level and fitting a smooth curve to those points. Figure 4.5 shows two questions with different difficulty.

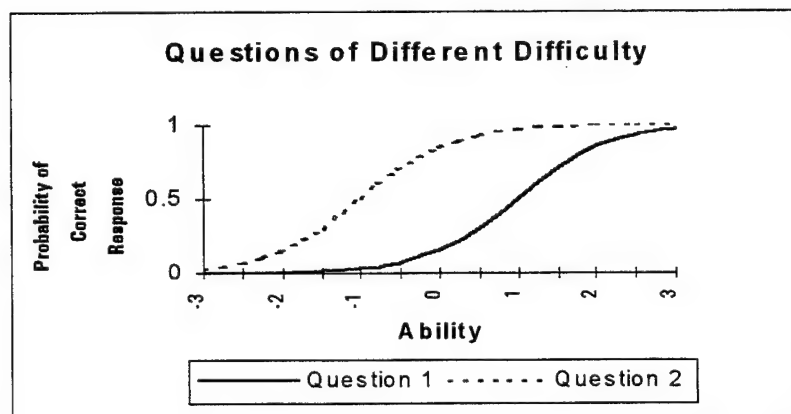
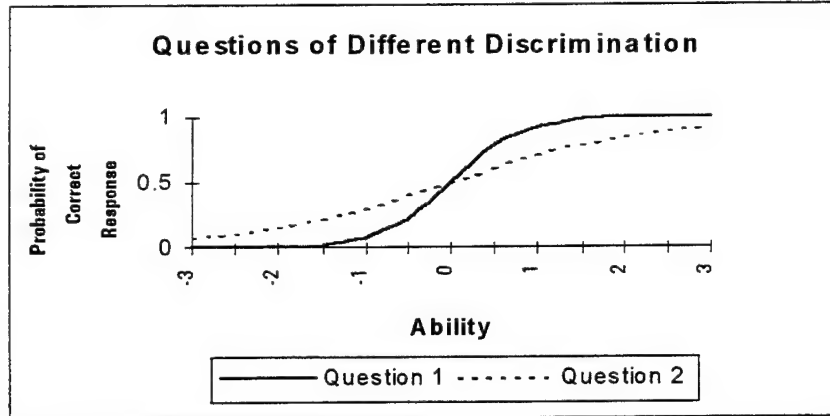


Figure 4.5. Item Characteristic Curve for two questions of different difficulty.

The question's difficulty is defined as the ability level where the probability of students of that ability getting a correct response is 0.5. For question 1 the difficulty is estimated to be 1 while for question 2 the difficulty is estimated at -1.0. The discrimination ability of



the question is proportional to the slope of the curve at the point where the probability of getting the question correct is 0.5. Figure 4.6 shows two ICCs with different discrimination abilities.



**Figure 4.6.** Item characteristic curves for two questions with different discrimination factors. Question 1 has better discrimination than question 2.

## 2. Types of Models

### a. One Parameter Logistic Model

The one parameter logistic model is one of the more widely used IRT models. The item characteristic curves for the one-parameter model is given by the equation

$$P_i(\theta) = \frac{e^{(\theta - b_i)}}{1 + e^{(\theta - b_i)}} \quad i = 1, 2, \dots, n \quad (4.1)$$

where

$P_i(\theta)$  is the probability that a randomly chosen student with ability

$\theta$  answers question  $i$  correctly,

$b_i$  is the question  $i$  difficulty parameter,

$n$  is the number of questions in the test.

The difficulty parameter  $b_i$  is the ability  $\theta$  where the probability of getting a correct response is 0.5. The higher the  $b_i$  value, the higher the ability required for the student to get the question correct. The difficulty parameter can vary from  $-\infty$  to  $+\infty$  depending on the scale used for ability but is usually between -3.0 to +3.0. A question with a difficulty parameter of 2.0 would be considered very difficult while one with that of -2.0 would be considered very easy. (Hambleton et al, 1991).

The one-parameter model assumes that all questions have the same discriminating value. There are no other item characteristics that define the question. In addition, there is no consideration that the student might guess at an answer, which would be possible on a multiple choice test.

#### **b. Two-Parameter Logistic Model**

To account for differences in the discriminating ability of questions, the two-parameter model was first developed by Lord (1952) and was based on the cumulative normal distribution. A similar and more commonly used model introduced by Birnbaum is to substitute the two parameter logistic function for the two parameter ogive function as the form of the ICC. (Hambleton et al, 1991) The item characteristic curves for the two-parameter logistic model are given by the equation

$$P_i(\theta) = \frac{e^{Da_i(\theta-b_i)}}{1 + e^{Da_i(\theta-b_i)}} \quad i = 1, 2, \dots, n. \quad (4.2)$$

In this model the parameters are defined as in the one-parameter model with new parameter  $a_i$  called the item discrimination factor and a scaling factor  $D$ . The item discrimination factor or parameter  $a_i$  is proportional to the slope of the ICC at the point  $b_i$  on the ability scale. The higher the value of  $a_i$ , the more discriminating the question. Questions with higher discrimination values are more useful for separating

students into different ability levels while a question with the discrimination factor near zero would provide no discrimination. The value of  $a_i$  is usually between 0 and +2.0.  $D$  is a scaling factor equal to 1.7 when the logistic model is used. Like the one parameter model, the two parameter model does not provide for guessing. (Hambleton et al, 1991)

### c. Three Parameter Model

The three-parameter model adds a factor that takes into account the possibility for the student to guess the correct answer. This model is most useful for tests with multiple choice items and is not used in this study since none of the tests are multiple choice. For a full discussion of the three-parameter model see Hambleton et al (1991).

## C. FITTING THE MODEL

### 1. Model Used

The model that is used for analyzing the tests is called the two-parameter model. In fact, there are more than two parameters to be estimated in this model since an ability for each student must be estimated along with the two parameters for each question. The probability of getting a correct response on question  $i$  for the two-parameter model given in Equation 4.2 is equivalent to the logistic regression model

$$\log\left(\frac{P_i}{1-P_i}\right) = Da_i(\theta - b_i). \quad (4.3)$$

where

$$P_i = P_i(\theta).$$

When student ability  $\theta$  is also unknown and considered a parameter to be estimated, the logit of the probability of a correct response (Equation 4.3) is not linear in the parameters. This model does not fall into the generalized linear model framework (eg McCullough and Nelder (1989)) and thus the usual packages for fitting logistic regression models are not directly applicable for estimating the question parameters  $a_i$  and  $b_i$  and ability  $\theta$ .

Techniques to estimate both the students' ability parameters and the question's parameters simultaneously are iterative. They require that an estimate be made of either the students' ability parameters or the question's parameters to start with. Both Hambleton(1983) and Baker (1992) suggest starting with an estimate of the student's ability parameters such as the normalized test scores. The normalized scores for all the students are used to estimate the question parameters and then these item parameters are used in an iterative process to estimate the student's ability. Methods for estimating ability and question parameters are discussed in the following sections. These steps are repeated until the change of parameters between each estimation of student's abilities and question parameters is within acceptable limits.

## **2. Item Parameter Estimation**

The parameter estimation for each question is made using the data, the estimate of each students ability, and the students score on each question ( 0 or 1). When  $\theta$  is known the logit in Equation 4.3 is linear in the remaining parameters making them easily estimated using a logistic regression. This generalized linear model is fit separately for each question giving a Maximum Likelihood Estimator (MLE) for  $a_i$  and  $b_i$ . Students' scores on a question are the response variables and their estimated ability is the prediction variable. Once obtained, these estimates are used in place of the question parameters and the student's abilities are estimated as described in the next section. This is repeated until convergence of all parameter estimates is obtained.

### 3. Student Ability Parameter Estimation

Although it is tempting to stop once the question parameters have been estimated, it is necessary to continue and estimate the ability parameters until the difference between iterations is below a satisfactory level. Baker (1992) derives the following formula using the first and second derivatives of the likelihood function to iteratively solve for the MLE's of the ability parameters. To distinguish between students a subscript  $j$  is added to ability, probability of correct response, and the response variable. Let  $J$  be the total number of students and,

$\theta_j, j=1, \dots, J$ , be student  $j$ 's ability,

$P_{ij}=P_i(\theta_j)$  be the probability of student  $j$  getting question  $i$  correct

and

$Q_{ij}=1-P_{ij}$ .

If  $[\hat{\theta}_j]_t$  is the estimate of  $\theta_j$  at the  $t^{th}$  iteration then

$$[\hat{\theta}_j]_{t+1} = [\hat{\theta}_j]_t + \left[ \frac{\sum_{i=1}^n Da_i (u_{ij} - P_{ij})}{\sum_{i=1}^n (Da_i)^2 P_{ij} Q_{ij}} \right]_t \quad (4.4)$$

where

$u_{ij} = 0$  for an incorrect response from student  $j$  on item  $i$   
 $1$  for a correct response from student  $j$  on item  $i$ .

Using the estimated question parameters and the previous estimate of ability  $[\hat{\theta}_j]_t$ ,  $a_i$ ,  $P_{ij}$ , and  $Q_{ij}$  are replaced on the right hand side of Equation 4.4 to give the new abilities  $[\hat{\theta}_j]_{t+1}$ . These new abilities are standardized and then used to give new estimates of the

question parameters. This procedure is repeated until the difference between the previous estimation of ability and question parameters are small enough.

#### **4. Preparing the Data**

When estimating parameters for a test it is necessary to remove questions and students that will cause problems for the model. Questions that are either missed by all students or correctly answered by all students lead to difficulty parameters that are infinite and cannot be estimated. In addition, the ability of students who answer all questions correctly or miss all questions will approach infinity or negative infinity. (Baker, 1992). Similar problems arise for questions that only a *few* students answer correctly or incorrectly and with students who either answer a *few* questions correctly or incorrectly. The data must be screened to remove these types of questions and students prior to parameter estimation. Additionally, questions with low discrimination or no discrimination will cause the difficulty of the question to approach infinity and must also be removed. Once the data is properly screened, the parameter estimation process can begin.

### **D. MODEL FIT**

#### **1. Screening Results**

The results of screening the Spanish Subskill tests are shown in Table 4.1. Spanish FLO's 30 and 40 require little screening of the data. Spanish FLO's 60, 70, and 90 require extensive screening which lead to poor results in the analysis. As the number of usable questions decreases it becomes more difficult to estimate student abilities. This in turn results in unstable estimates of question parameters. The end result is that the algorithm fails to converge and parameters cannot be estimated. Finally, after screening

Test	Total Number of Questions	Number of Questions all Correct	Number Less Than 4 Wrong	Number Less Than 4 Right	Number of Questions With No Discrimination	Questions Remaining
FLO30	30	0	0	2	0	28
FLO40	32	0	4	0	1	27
FLO60	40	2	22	0	10	6
FLO70	30	3	9	0	9	9
FLO80	30	9	12	0	7	2
FLO90	32	0	10	1	11	10

**Table 4.1.** Results of data screening for the Spanish Subskill test data.

Spanish FLO 80 there was not enough data remaining to conduct an analysis.

Since there were only 29 students in the Russian data set, the Subskill test data set was not large enough to conduct IRT analysis.

## 2. Results of IRT Parameter Estimation

As can be expected from the results of the screening, Spanish FLO's 30 and 40 have the best results, while FLO's 60 and 70 have the worst. Table 4.2 shows how many

Test	Questions Remaining after screening	Questions Successfully Modeled
FLO 30	28	25
FLO 40	27	27
FLO 60	6	2
FLO 70	9	5
FLO 80	2	0
FLO 90	10	6

**Table 4.2.** Number of questions successfully modeled for each Spanish FLO.

questions are successfully modeled using the IRT parameter estimation technique described. The estimated values of the parameters for each test can be found in Appendix C.

### 3. Goodness of Fit

The item response model given in Equation 4.3 is a first approximation to the relationship between a student's ability and the probability of answering the question correctly. Very little work has been done on methods for checking the adequacy of an item response model. A recent paper suggests fitting a nonparametric item response model such as a generalized additive model (Douglas, 1995) and then checking to see how close the parametric ICC is to the nonparametric ICC. For Spanish FLO 30 and 40, a generalized additive model is fit to each question separately treating the estimated ability from the item response model as the explanatory variable, i.e.

$$\log\left(\frac{P_i(\theta)}{1 - P_i(\theta)}\right) = \alpha_i + B_i s(\theta) \quad (4.5)$$

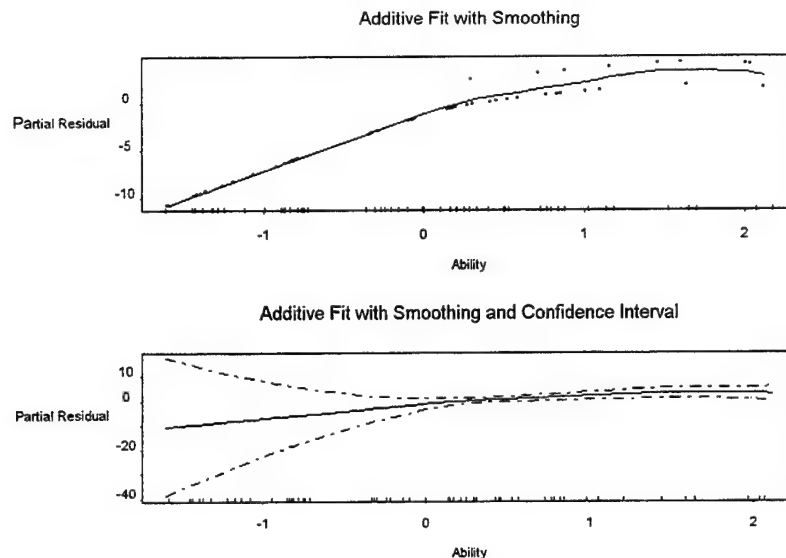
where  $s(\theta)$  is a "smooth" function of  $\theta$  to be estimated in the fit of the generalized additive model. Rather than follow Douglas' suggestion to plot the parametric ICC versus the nonparametric ICC, it is often more revealing to look at some form of residuals, when examining the fit of a model. Plots of partial residuals versus ability (Hastie and Tibshirani, 1991) are used to assess fit. The partial residuals are given by

$$\frac{u_{ij} - \hat{P}_{ij}}{\hat{P}_{ij}(1 - \hat{P}_{ij})} + \hat{s}_i(\theta_j) \quad (4.6)$$

where  $\hat{P}_{ij}$  is the estimate of  $P_{ij}$  computed from the generalized additive model. A linear relationship in the plot of the partial residuals versus ability indicates that  $s(\theta)$  is indeed linear and the IRT model is adequate. A nonlinear trend indicates that the IRT model probably does not adequately describe the relationship between students ability and the probability of answering correctly.



From the plots of both the Spanish FLO's 30 and 40 question it appears that about 80% of the questions behave in a linear fashion. However, the remainder of the questions show that the relationship between ability and logit of the probability of answering the question correctly is not linear. Figure 4.7 shows a plot of the partial residuals of a



**Figure 4.7.** Plot of a the partial residuals from a general additive model of a question from Spanish FLO 30 showing the lack of a linear relationship.

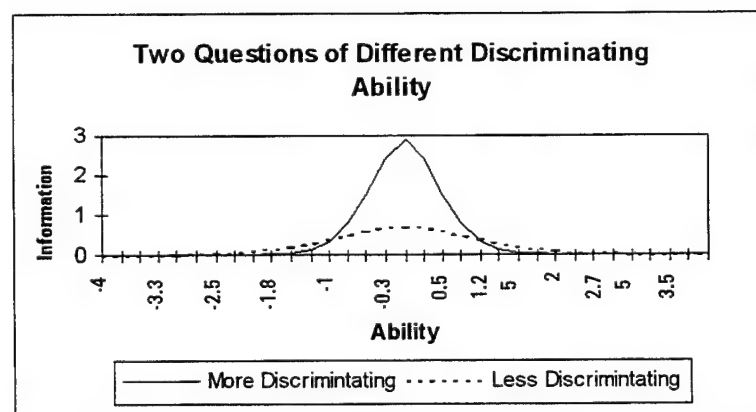
smoothed general additive model. From this figure it can be seen that the relationship is not linear. After examining the width of the confidence interval band in the lower graph of Figure 4.7 it is apparent that care must be taken in evaluating the residual plots. For many questions in FLO's 30 and 40 the confidence band is wide enough that it is not possible to tell whether the trend in the partial residual plot is an artifact of variation in the data or an indication that the IRT model does not fit. One reason for the large confidence bands are the small sample sizes. With a larger sample size, the confidence band will tighten and the relationship between  $\theta$  and  $P_i(\theta)$  or the logit will be more clear. "The linear model is a convenient but crude first-order approximation to the prediction surface, and in many cases it is adequate" (Hastie and Tibshirani, 1991). Thus, with the sample size available

and for the purposes of this look at the discrimination ability of the FLO's, the IRT model is adequate.

## E. INTERPRETING THE MODEL

### 1. Information Functions

One of the more useful aspects of IRT is the information function. When the question parameters are known, the amount of information provided by a question at a specific ability level can be determined using a question's information function. The information functions from all the questions in a test can be combined to determine the amount of information provided by the test at each ability level. This can be used to construct or modify a test to ensure that a certain cutoff ability level is tested for or to that a wide range of ability is covered by the test. The amount of information provided by a question is directly related to the discrimination power of that question. Figure 4.8 is a



**Figure 4.8.** Information functions of two questions of different discrimination parameter but the same difficulty parameter. The information values are inversely related to the  $SE(\theta)$ .

graph of two questions' information functions, each question with the same difficulty but different discrimination parameters. The peak of each function occurs at the same point

on the x-axis (ability), but the question with poorer discrimination does not provide as much information. The item information function is given by the equation

$$I_i(\theta) = \frac{[P'_i(\theta)]^2}{P_i(\theta)Q_i(\theta)} \quad i = 1, 2, \dots, n. \quad (4.7)$$

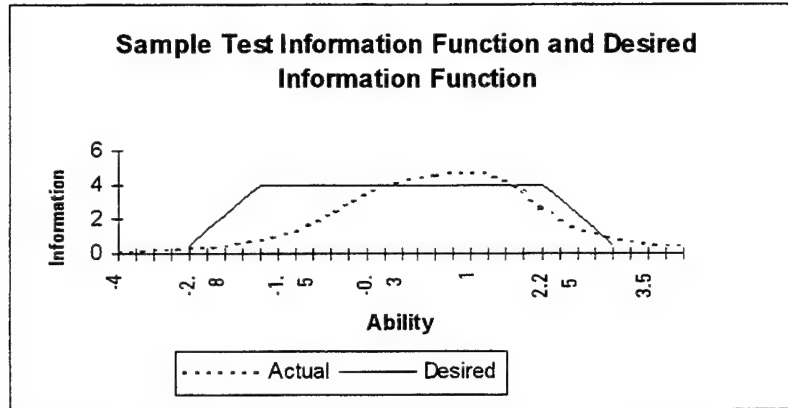
By summing the information functions over questions in a test, the test information function is given by

$$I(\theta) = \sum_{i=1}^n I_i(\theta). \quad (4.8)$$

The amount of information provided by a test at  $\theta$  is inversely related to the precision with which the ability is estimated at that point (Hambleton et al, 1991). Using the relationship

$$SE(\hat{\theta}) = \frac{1}{\sqrt{I(\theta)}} \quad (4.9)$$

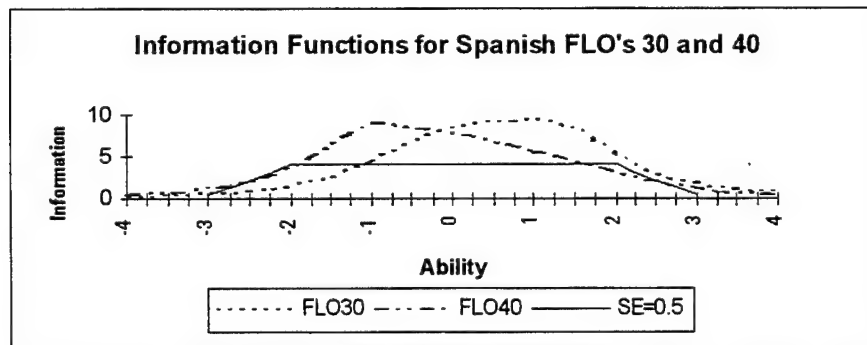
where  $SE(\hat{\theta})$  is the standard error of  $\hat{\theta}$ , it is possible to determine the information required for a specific precision level. By comparing plots of the desired information level to the actual information provided by a test it is possible to determine which types of questions need to be added to the test to achieve that level. Figure 4.9 shows the graph of a sample test's information function and the information function with SE equal to 0.5. From this figure it can be determined that both easier and more difficult questions need to be added to achieve the precision level of 0.5 over an ability range of -2.0 to +2.0.



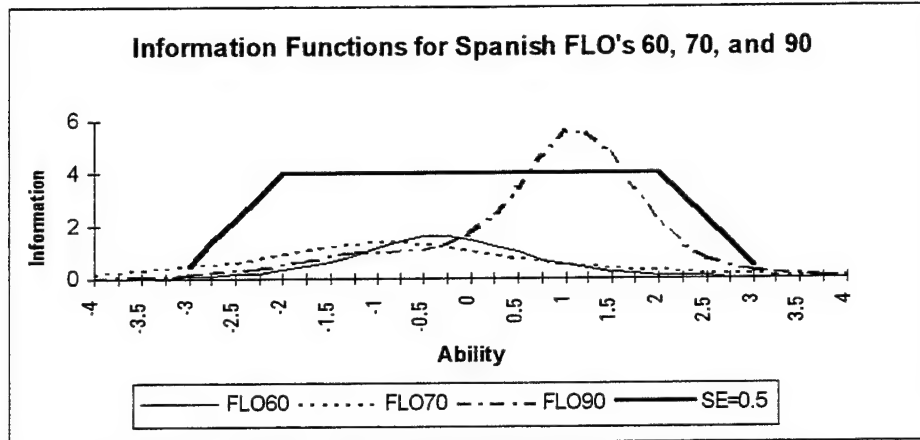
**Figure 4.9.** Graph of desired information function and sample test information function. The information values are inversely related to the  $SE(\theta)$ .

## 2. Actual Information Functions

Figure 4.10 shows the information functions for both Spanish FLO's 30 and 40 with an additional curve representing desired precision level of  $SE=0.5$  for comparison. Figure 4.11 shows the information functions for Spanish FLO's 60, 70, and 90 also with desired curve representing precision levels of  $SE=0.5$ .



**Figure 4.10.** Information functions for Spanish FLO's 30 and 40 with desired precision curves of  $SE=0.5$ . The information values are inversely related to the  $SE(\theta)$ .



**Figure 4.11.** Graphs of the information functions for Spanish FLO's 60, 70, and 90 with desired curve representing  $SE=0.5$ . The information values are inversely related to the  $SE(\theta)$ .

## F. DISCUSSION OF RESULTS

The results of the data screening reveal that for Spanish FLO's 60, 70, 80 and 90, many of the questions are not providing information to differentiate between students. As seen in Table 4.1, the majority of the questions are either answered correctly by all the students or provide no discrimination between students of different ability. The few questions that do remain provided limited information about the student's ability as can be seen in Figure 4.11. From these results, many of the questions on these tests can be replaced with more difficult and better discriminating questions to provide precise information about the student's ability. One exception to this may be with Spanish FLO 60 which tests number transcription ability at a specific rate. For this type of test it may be difficult to vary the difficulty of the questions. Since the students at DLI are heavily exposed to numbers throughout their training, they may all be able to transcribe at the desired rate. The questions that show poor discrimination should still be examined to determine if they should be removed.

The two tests that modeled well are Spanish FLO 30 and 40. The questions on these tests covered a broad range of ability and most have sufficient discrimination ability.

of a precision level equivalent to a SE of 0.5 is desired it would be necessary to add questions in the -2.5 to the -1.0 difficulty level for FLO 30.

When discussing the results of IRT analysis, it must be remembered that it is assumed that the data used is representative of the population that takes the test. If DLI plans on using the Subskill tests to evaluate field personnel as it does with the DLPT, it will be necessary to have field personnel take the tests and evaluate their individual question responses. The inclusion of field personnel should decrease the difficulty level of the questions since adding the field personnel will change the population and should result in a higher test average for the tests. For questions that the current population of test takers found difficult the difficulty will be lowered on a standardized scale.

In general, the IRT analysis shows that many of the questions on the Subskill tests can be eliminated without affecting the determination of the student's ability. Using IRT, it is possible to determine how many of these question need to be replaced with more difficult and better discriminating questions to achieve a desired precision level.



## **V. TEST REDUCTION**

### **A. MOTIVATION**

#### **1. Background**

Just prior to graduation students at DLI take the 10 Subskill tests and 3 Defense Language Proficiency Tests (DLPTs). As mentioned previously, 6 of the 10 Subskill tests are graded at the Test Management Center by any one of the three GS-5 employees whose job is to grade these tests. The remaining 4 Subskill tests are graded at the respective language school. Each rater can grade an average of 7 tests an hour, requiring about 26 hours to grade all six of the Subskill tests for a class of size 30. Since both the DLPTs and the Subskill tests determines the students' ability to read, speak and listen to the target language, it is possible that the tests will measure overlapping abilities and that at least one of the tests may be eliminated. A reduction in the number of Subskill tests would result in a monetary savings by reducing the workload for the raters, test proctors, and students. The sponsors of the Subskill tests are not willing to reduce the number of Subskill tests unless it can be shown that no significant loss of information about the students' ability to perform the Final Learning Objectives will occur. The DLPTs are not being considered for elimination since they are used to evaluate both students and field personnel proficiency levels.

#### **2. Methods to be Employed**

Principle components and a second method based on multiple correlation both presented by Jolliffe (1986) are used to determine which Subskill tests can be removed. The goal of both methods is to eliminate only those tests that reduce the amount of information in the tests by a small amount. In other words, remove those tests which are



highly collinear with a linear combination of the remaining tests. Because the two methods use different criteria to select tests to remove, they may select tests in a different order and sometimes different tests altogether.

### **3. Data Examined**

To continue with the analysis of the previous two chapters, these methods are applied to Spanish and Russians students graduating between October 1994 and March 1996. Variability in test scores can be caused by differences in ability and by tests with poorly discriminating questions. From Chapter 4 it was shown that Spanish FLO 60 had 10 poorly discriminating questions. To determine if removing the 10 poorly discriminating questions will affect the order or selection of tests to be removed and the amount of variance still explained by the remaining tests, two additional data sets are studied. The two data sets are constructed using the scores of the Spanish students studied in the previous two chapters. The first data set is constructed from the data as retrieved from DLI's data base. The second data set is the same as the previous with the exception of the FLO 60 data. The 10 poorly discriminating questions are removed from the test and the students' grades are then recomputed.

## **B. METHODOLOGY**

### **1. The Principal Component Method**

Principle component analysis is a common method for reducing the dimensions of a data set while retaining as much as possible of the variation in the original data set. "The reduction is achieved by transforming to a new set of variables, the principle components, which are uncorrelated, and which are ordered so that the first few retain most of the variation present in all of the original variables" (Jolliffe, 1986). One problem with using

principle components to reduce the dimensions of a data set is that they are often linear combinations of all of the variables (Dunteman, 1983). This would defeat the purpose of reducing the dimensions of the data set. However, principle components can still be used to determine which variables to remove from the data set. In a method referred to as the B2 model a principle component analysis is performed on all the original  $K$  variables and the eigenvalues are inspected. For all principle components whose eigenvalues are less than some  $\lambda_0$ , the corresponding eigenvectors are inspected starting with the vector with the smallest eigenvalue and continuing with the vector having the next smallest until  $\lambda_0$  is reached. For each vector, the variable associated with the largest component of the vector is removed from the data set. If the variable has already been removed, the variable associated with the next largest component is removed. Jolliffe (1972) recommends a value for  $\lambda_0$  of 0.7. This value leads to a data set that only retains 60-70 % of the original data set's variance, which is not an acceptable amount. Therefore, in this analysis variables are removed until less than 90% of the original variance is still explained by the remaining variables. (Jolliffe, 1986)

## 2. The Multiple Correlation Method

A closely related method to remove variables from a data set uses multiple correlation. This method explained by Jolliffe (1972) uses multiple correlation to pick  $p$  variables to describe the variance of the original  $K$  variables. "Method A2 is a step-wise method which first rejects that variable which has a maximum multiple correlation with the remaining  $K-1$  variables. Then at each stage, when  $q$  variables remain, the variable having the largest multiple correlation with the other  $q-1$  variables is rejected." The process continues until  $p$  variables remain. For this analysis, the process continues until the less than 90% of the original variance is explained by the remaining  $p$  variables.

### 3. Comparing Methods

To determine the effectiveness of each method it is necessary to measure the amount of variance remaining after the removal of a test from the data set. Jolliffe (1973) uses the following formula to compute the remaining variance in a test:

$$= \frac{p + \sum_{i=p+1}^K r_i}{K} \quad (5.1)$$

where

$p$  = the number of remaining tests or variables

$r_i$  = the multiple correlation factor of removed test  $i$  and the remaining tests

$K$  = the original number of tests.

### C. RESULTS

The A2 and B2 methods were first applied to all Spanish and Russian students graduating between October 1994 and March 1996. The results for the Spanish students are shown in Table 5.1 and for the Russian students in Table 5.2. There were originally 529 Spanish students and 349 Russian students in the data base. Due to missing test scores in the data base, the data set was reduced to 426 Spanish and 262 Russian students. The results for the Spanish data set show that both methods remove FLO 40 first followed by FLO 30. Removing any more of the Subskill tests reduces the explained variance to below 90 %, which was the desired cutoff level. The results for the Russian data set are different than that of the Spanish but also indicate that up to two Subskill tests can be removed without going below the desired 90 % explained variance. In addition the methods suggest a different order to remove the tests but pick the same two to be removed when removing two tests.

Number of Tests Removed	Method	Tests Removed	Variance Explained by the Remaining Tests (in percent)
1	Multiple Correlation	FLO 40	96.45
	Principle Components	FLO 40	96.45
2	Multiple Correlation	FLO 40, FLO 30	91.57
	Principle Components	FLO 40, FLO 30	91.57
3	Multiple Correlation	FLO 40, FLO 30, FLO 50	86.40
	Principle Components	FLO 40, FLO 30, FLO 50	86.40

**Table 5.1.** Results of variable reduction using data on Spanish students graduating between October 1994 and March 1996.

Number of Tests Removed	Method	Tests Removed	Variance Explained by the Remaining Tests (in percent)
1	Multiple Correlation	FLO 30	95.54
	Principle Components	FLO 10	95.38
2	Multiple Correlation	FLO 30, FLO 10	90.86
	Principle Components	FLO 10, FLO 30	90.86
3	Multiple Correlation	FLO 30, FLO 10, FLO 40	85.80
	Principle Components	FLO 10, FLO 30, FLO 50	85.27

**Table 5.2.** Results of variable reduction using data on Russian students graduating between October 1994 and March 1996.

Next, the methods were applied to a smaller subset of the Spanish students to determine if removing the 10 poorly discriminating questions from the Spanish FLO 60 would affect the results. Table 5.3 shows the results of applying both methods to the subset without changing the FLO 60. Table 5.4 shows the results of both methods with the 10 questions removed. For both sets of data it is possible to remove three Subskill tests without losing more than 10 % of the variance. For the unaltered FLO 60 data set,

Number of Tests Removed	Method	Tests Removed	Variance Explained by the Remaining Tests (in percent)
1	Multiple Correlation	FLO 40	98.27
	Principle Components	FLO 30	97.97
2	Multiple Correlation	FLO 40, FLO 70	95.09
	Principle Components	FLO 30, FLO 40	93.89
3	Multiple Correlation	FLO 40, FLO 70, FLO 30	90.52
	Principle Components	FLO 30, FLO 40, FLO 70	90.52
4	Multiple Correlation	FLO 40, FLO 70, FLO 30, FLO 90	86.97
	Principle Components	FLO 30, FLO 40, FLO 70, FLO 90	86.97

**Table 5.3.** Results of variable reduction using data from DLI database on Spanish students studied in the previous two chapters.

Number of Tests Removed	Method	Tests Removed	Variance Explained by the Remaining Tests (in percent)
1	Multiple Correlation	FLO 40	97.95
	Principle Components	FLO 30	97.92
2	Multiple Correlation	FLO 40, FLO 30	95.47
	Principle Components	FLO 30, FLO 40	95.47
3	Multiple Correlation	FLO 40, FLO 30, FLO 90	92.37
	Principle Components	FLO 30, FLO 40, FLO 70	91.76
4	Multiple Correlation	FLO 40, FLO 30, FLO 90, FLO 70	88.30
	Principle Components	FLO 30, FLO 40, FLO 70, FLO 90	88.30

**Table 5.4.** Results of variable reduction using data from DLI database on Spanish students studied in the previous two chapters with the FLO 60 modified by removing the 10 poorly discriminating questions.

the methods select the same three variables but in a different order as did the Russian.

Finally, using the modified data set, both methods select FLO's 30 and 40 as the first two tests to be removed but then select different tests to be removed after that.

## D. DISCUSSION OF RESULTS

The three features to examine in the results of the variable reduction methods are the difference between the methods, the effects of removing the 10 questions from Spanish FLO 60, and the number of tests that can be removed.

The results from Table 5.1 show that there is no difference in methods when examining the larger data base of Spanish students. When applied to the Russian data set, as well as to the smaller Spanish data set, the methods select the same two Subskill tests, but in a different order. In addition, the smaller Spanish data sets allow for the removal of a third test. This third test is the same for the unmodified data set but different for the modified data set. The reason that the methods select tests in a different order and sometimes different tests altogether is that the A2 method (multiple correlation method) considers the impact of removing a test on the remaining data set. The B2 method (principle components method) focuses on removing the test that explains the most variance in that principle component without considering how much that test contributes to the variance in the remaining principle components. Because of this, the B2 method, when it differs from the A2 method, will not always select the best test to remove. Jolliffe (1972) also found that the A2 method was the better of the two methods but felt that neither method was notably better or worse than the other for artificial data.

The next item to examine is the effect of removing the 10 questions from the Spanish FLO 60 in the small data set. Recall that the purpose of removing the 10 poorly discriminating questions was to see if, by removing some of the variability in that test if the order or selection of tests to be removed and the amount of variance still explained by the remaining tests would be affected. As seen in Tables 5.3 and 5.4 there was little effect on the amount of variance explained after the same number of tests had been removed. There was a difference in selection of tests to be removed using the A2 method. FLO 70 is selected to removed second followed by FLO 30 third in the unmodified data set whereas FLO 30 is selected second followed by FLO 90 third when examining the modified data

set. The end result is that there is less than a 1 percent difference in explained variability caused by removing the 10 questions. This indicates that the 10 questions do not account for a large amount of the variance present in the data set.

Finally, by examining the results of all four data set it can be seen that at least one of the Subskill tests can be removed from both languages. FLO 40 could be removed from the Spanish and FLO 30 could be removed from the Russian test batteries with no significant loss of information about the students.

## **VI. SUMMARY/RECOMMENDATIONS**

### **A. SUMMARY**

The Defense Language Institute developed the Subskill tests to determine if the graduating students have met the specific training objectives as outlined by the National Security Agency and the Defense Intelligence Agency. These Subskill tests have been in place for over two years and it is now possible to evaluate them and determine if any changes should be made to improve their consistency and efficiency.

For the grades to be consistent, a student should receive the same score independent of who grades their test. The analysis of variance shows that assignment of scores is not consistent between raters for the majority of Subskill tests in Russian and Spanish. In particular the Spanish FLO's 30 and 90 have the largest rater magnitudes allowing a difference in grades as large as 16 points.

An efficient test is made of questions that discriminate between students of differing abilities and are of different difficulty levels. The Spanish FLO's 30 and 40 tests are made of questions that do this. In contrast are the Spanish FLO's 60, 70, 80 and 90 which consist of many questions that are either too easy for all the students or do not discriminate among students of different ability. The questions that do not discriminate create unnecessary variance in the data set and provide no information about the student's ability.

Finally, although there was surprisingly little redundancy between the Subskill tests and the DLPTs, the Subskill tests and the DLPTs provide some redundant information about the student's abilities. The Spanish FLO 40 and Russian FLO 30 can be removed with less than a 5% loss of variability in the data set.



## **B. RECOMMENDATIONS**

Each of the above findings should be addressed to ensure that the Subskill tests provide consistent, efficient data to be used to compare students and determine if the students have met the training objectives.

The rater effect can be eliminated by reducing the amount of subjective evaluation of the student's responses and allowing the rater to give partial credit. By providing the rater with an English translation of the material and questions given to the student the rater will more accurately evaluate if the student has understood the question and responded correctly. In addition by allowing partial credit, the differences between rater evaluations will be smoothed.

More data on the Spanish FLO's 60, 70, 80 and 90 should be gathered to determine which test questions provide the same information about the student's ability. To do this it will be necessary to record the students' scores on individual questions. The raters currently fill out a computer scan sheet that reads the score for individual questions but this information is not recorded or maintained. The Test Management Center should maintain this information for future use. Many of the questions on the Spanish FLO's 60, 70, 80, and 90 were answered correctly by all students taking the tests. These questions should be replaced by questions of greater difficulty so as to provide more information about the student's ability. The questions that do not discriminate should be fixed or eliminated to reduce the amount of variance in the tests.

The number of Subskill tests can be reduced in both Russian and Spanish. This will reduce the amount of time spent administering and grading the tests as well as reduce the amount of information that needs to be maintained in the database. Consideration must be given to the effects of test removal on teaching methods and students' motivation to ensure that students continue obtain the desired language abilities.

The rater effect issue should be addressed first. By reducing the subjective evaluation of the student's responses, the relationship between the student's ability and

response to a question may change. This could effect the discrimination power of a test question which would in turn result in the question providing useful information about the student's ability. By removing the rater effect the variance in the data base may also be reduced which could change both the order in which tests are selected to be removed and the amount of variance explained by the remaining tests.

These changes will ensure that the Subskill tests developed by the Defense Language Institute will continue to efficiently provide useful consistent information about the students' language abilities.



## APPENDIX A. RATER EFFECTS SUMMARY TABLE

Test	Spanish			Russian		
	Grader 1	Grader 2	Grader 3	Grader 1	Grader 2	Grader 3
FLO 30	-4.3894	10.4365	-6.0516	-8.161	2.069	6.092
FLO 40	-1.726	-0.1786	1.905	-2.797	1.686	1.111
FLO 60	0.0298	-0.1488	0.1190	-0.02874	0.4023	-0.3736
FLO 70	-6.071	2.143	3.929	-1.341	-0.3065	1.6475
FLO 80	0.3273	0.6182	-0.9455	1.1111	-4.2912	3.1801
FLO 90	7.9382	0.4490	-8.3872	3.125	-0.1078	-3.0172

**Table A-1.** Summary Table of the rater effects (points on a 100 point test) for both Spanish and Russian Subskill tests.

Test	Spanish			Russian		
	Grader 1	Grader 2	Grader 3	Grader 1	Grader 2	Grader 3
FLO 30	4.072407	4.257052	4.855999	5.205443	4.481238	6.039411
FLO 40	3.461636	3.378495	3.716225	3.798504	4.558271	3.72086
FLO 60	0.9529895	0.8333706	0.9417565	1.229651	1.366055	1.775972
FLO 70	4.310946	3.730607	3.858146	4.04319	3.169194	3.137371
FLO 80	0.7869588	0.8573562	0.963506	2.672612	2.072839	3.262801
FLO 90	4.405731	3.409063	4.266555	4.148024	4.132532	5.953696

**Table A-2.** Summary Table of the standard deviation of the rater effects for both Spanish and Russian Subskill tests.



## APPENDIX B. RESULTS OF IRT ANALYSIS

Test	FLO 30		FLO 40	
Question	Discrimination	Difficulty	Discrimination	Difficulty
1			0.34209	0.60711
2	0.65204	0.81167	0.76504	-1.2717
3	0.76636	0.33544		
4	1.82756	1.46266	0.6502	0.61854
5	0.49392	5.13756	1.47022	0.03907
6	1.35672	0.91748	0.14762	-2.9153
7	0.70393	2.59003		
8			0.79932	0.95586
9	0.30809	1.29823	1.08681	-0.4285
10	1.42263	1.33465	0.96512	0.70131
11	0.25755	1.72006	1.00159	1.40624
12	0.23114	-1.8944	1.00336	-0.185
13	0.7918	0.03094	0.47112	0.90085
14	0.65978	0.53889		
15	0.19879	3.55872	0.31207	0.36201
16	0.76315	0.4139	0.42472	-0.1745
17	0.52602	-0.5328	0.22919	0.09372
18				
19	1.52085	-0.5827	0.80814	1.57112
20	1.44141	-0.1557		
21			0.71779	-1.8062
22	0.87481	-1.6094	1.21803	1.64727
23	0.58767	-0.1353	0.39674	0.78116
24	0.3892	2.8992	1.10136	-1.6458
25			0.53417	-0.2441
26	0.25716	-1.1477	0.36638	0.31357
27	0.7535	0.91575	1.08834	-1.0517
28	1.33622	0.38368	0.89291	-0.9072
29	1.0466	1.89342	0.80613	-0.1257
30	1.33622	0.38368	0.98729	1.11133
31			0.52382	0.9355
32			0.0678	3.14E-01

**Table B-1.** Question difficulty and discrimination parameters for Spanish FLO's 30 and 40 found using IRT analysis.

Test	FLO 60		FLO 70		FLO 90	
Question	Discrimination	Difficulty	Discrimination	Difficulty	Discrimination	Difficulty
1						
2						
3					0.997437	-0.73496
4						
5						
6	0.99912	-0.1385				
7						
8					0.531537	-1.24748
9						
10						
11						
12						
13			0.724721	-1.16264	1.927979	-1.10155
14			0.387	-0.47342		
15						
16						
17						
18						
19			0.7664	-0.83737		
20						
21						
22						
23						
24						
25					0.634196	1.103855
26					0.891643	-0.66949
27	1.13581	-0.4661				
28			0.487187	-1.68542		
29					0.808497	-0.70838
30			0.450146	0.508869		
31		2.19982				
32						
33						
34						
35						
36	1.5817	-0.1355				
37	1.13149	-0.4671				
38						
39						
40	0.18522	-1.0047				

**Table B-2.** Question difficulty and discrimination parameters for Spanish FLO's 60, 70 and 90 found using IRT analysis.

## LIST OF REFERENCES

- Baker, F. B., *Item Response Theory*, Dekker, 1992.
- DLI, *Final Learning Objectives for Basic Level Language Programs in the Defense Foreign Language Program*, 1995.
- Devore, J. L., *Probability and Statistics for Engineering and the Sciences 4th edition*, Wadsworth, 1995.
- Douglas, J., Nonparametric ICC Estimates To Assess Fit of Parametric Models, *Submitted for publication*, 1995.
- Dunteman, G. H., *Principle Components Analysis*, Sage Publications, 1983.
- Gibbons, J. D., *Nonparametric Statistical Inference*, Dekker, 1985.
- Hambleton, R. K., *Applications of Item Response Theory*, Educational Research Institute of British Columbia, 1983.
- Hambleton, R. K., Swaminathan, H., Rogers, H. J., *Fundamentals of Item Response Theory*, Sage Publications, 1991.
- Hastie, T. S. and Tibshirani, R. J., *Generalized Additive Models*, Chapman & Hall, 1991.
- Jolliffe, I. T., Discarding Variables in a Principle Component Analysis. I. Artificial Data, *Applied Statistics*, vol 21, 1972.
- Jolliffe, I. T., Discarding Variables in a Principle Component Analysis. II. Real Data, *Applied Statistics*, vol 22, 1973.
- Jolliffe, I. T., *Principal Component Analysis*, Springer-Verlag, 1986.
- Lindman, H. R., *Analysis of Variance in Experimental Design*, Springer-Verlag, 1992.
- Longford, N. T., Reliability of Essay Rating and Score Adjustment, *Program Statistics Research*, Technical Report No. 93-36, 1993.
- Lord, T. M., A Theory of Test Scores, *Psychometric Monograph*, No 7, Iowa City, 1952.



McCullough, P. and Nelder, J. A., *Generalized Linear Models* (2nd ed), Chapman & Hall, 1989.

Raymond, M. R., Houston, W. M., Detecting and Correcting for Rater Effects in Performance Assessment, *ACT Research Report Series*, 90-14, 1990.

## INITIAL DISTRIBUTION LIST

	No. Copies
1. Defense Technical Information Center 8725 John J. Kingman Rd., STE 0944 Ft Belvoir, Virginia 22060-6218	2
2. Dudley Knox Library Naval Postgraduate School 411 Dyer Rd. Monterey CA 93943-5101	2
3. Gordon Jackson Research and Analysis Division Defense Language Institute Presidio of Monterey CA 93944-5006	2
4. Professor Lyn D. Whitaker, Code OR/Wh Department of Operations Research Naval Postgraduate School Monterey CA 93943-5002	2
5. Professor Robert R. Read, Code OR/Re Department of Operations Research Naval Postgraduate School Monterey CA 93943-5002	2
6. Lieutenant Carlton L. Lavinder III USN 1127 E. Bay Shore DR Va Beach VA 23451	2